Super-SCOTUS: A multi-sourced dataset for the Supreme Court of the US

Biaoyan Fang[◇*] **Trevor Cohn**[◇†] **Timothy Baldwin**[◇,♠] **Lea Frermann**[◇]
[◇]The University of Melbourne, Australia
[♠]MBZUAI
{biaoyan,tcohn,tbaldwin,lfrermann}@unimelb.edu.au

Abstract

Given the complexity of the judiciary in the US Supreme Court, various procedures, along with various resources, contribute to the court system. However, most research focuses on a limited set of resources, e.g., court opinions or oral arguments, for analyzing a specific perspective in court, e.g., partisanship or voting. To gain a fuller understanding of these perspectives in the legal system of the US Supreme Court, a more comprehensive dataset, connecting different sources in different phases of the court procedure, is needed. To address this gap, we present a multi-sourced dataset for the Supreme Court, comprising court resources from different procedural phases, connecting language documents with extensive metadata. We showcase its utility through a case study on how different court documents reveal the decision direction (conservative vs. liberal) of the cases. We analyze performance differences across three protected attributes, indicating that different court resources encode different biases, and reinforcing that considering various resources provides a fuller picture of the court procedures. We further discuss how our dataset can contribute to future research directions.[1]

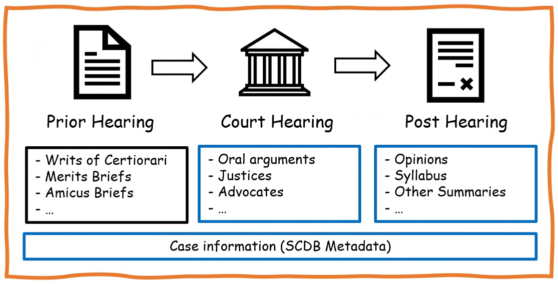


Figure 1: The jurisdiction procedure of the Supreme Court. Data resources highlighted in blue are included in Super-SCOTUS, our multi-sourced dataset.

1 Introduction

With the increasing attention to legal text processing, recent research has proposed various legal corpora, covering different sources, e.g., court documents (Zheng et al., 2021; Chalkidis et al., 2022a; Niklaus et al., 2021; Henderson et al., 2022; Locke and Zuccon, 2018) or legal contracts (Tuggener et al., 2020; Hendrycks et al., 2021; Lippi et al., 2019), and tackling legal analysis with a diverse set of tasks, e.g., text classification (Chalkidis et al., 2022a,b), juridical output prediction (Zhong et al., 2018; Cui et al., 2022) and case summarization (Ye et al., 2018; Shukla et al., 2022).

In the Supreme Court of the US (SCOTUS), several procedures contribute to the court results and reasoning process (Stern and Gressman, 1950). As illustrated in Figure 1, for instance, a *writ of certiorari* is needed for the petition of the appeal to the SCOTUS. *Merits briefs* in which petitioners and respondents lay out their arguments to the Court, are required before the oral arguments take place, which in turn can clarify or elaborate on points made in the briefs. Justices further discuss and decide the case, and afterwards issue *Opinions* which explain the reasoning behind the final judgment. We present a comprehensive data set of SCOTUS proceedings and meta data to enable holistic language analysis of the court.

Various SCOTUS resources have been gathered from different phases of the ruling procedure and formulated in diverse tasks in order to analyze different perspectives of the court, e.g., voting (Ruger et al., 2004; Katz et al., 2017; Dietrich et al., 2019), partisanship (Bergam et al., 2022) and topic prediction (Chalkidis et al., 2022a,b). However, each work comprises only a limited set of resources with a specific focus on one particular court perspective.

In light of this observation, we take a step forward and create a multi-sourced SCOTUS dataset,

[*]Now at CSIRO Data61.
[†]Now at Google DeepMind.
[1]Dataset and code are made available at https://github.com/biaoyanf/Super-SCOTUS.

Proceedings of the Natural Legal Language Processing Workshop 2023, pages 202–214
December 7, 2023 ©2023 Association for Computational Linguistics

i,e., Super-SCOTUS (Section 3). Figure 1 illustrates the scope and structure of our dataset. We focus on the phases of the court hearing and post hearing, augmented with relevant metadata on the cases and justices.[2] Our corpus connects publicly-available resources including oral arguments and various post-hearing annotations and summaries, including Opinions and case summaries.

To showcase the utility of our data set, we present a case study in Section 4 where we ask *how do different court documents reflect the direction of the final court decision (as conservative or liberal)*. We predict the decision direction from input documents from the phases of the court hearing (Oral arguments) and post-hearing (Syllabus). We consider three sensitive case attributes, derived from the meta data of our corpus, to analyse group disparity of our classifiers with respect to (1) Issue Area, (2) Vote Distribution, and (3) Winning Side. We observe that, compared to oral arguments, case direction is easier to decode from the syllabus. The group disparity across three protected attributes indicates that different court materials might encode different biases, suggesting the necessity of considering various resources for the full analysis of the Supreme Court. We finally point out potential directions and use cases for future work (Section 5).

In summary, our contributions are (1) a publicly available multi-sourced SCOTUS dataset, comprising various court information from different procedure phases. (2) a case study, addressing the research question of *how do the court materials reflect the case direction*, overall and wrt. discrepancy for three protected attributes (3) a summary of the different layers of legal texts in the Supreme Court and a discussion for future research directions building on our dataset.

2 Related Work

The judiciary in court, as the base of the legal system, has been one of the key focuses in the legal domain (Martin and Quinn, 2002; Epstein et al., 2010; Devins and Baum, 2017; Zheng et al., 2021; Fang et al., 2023a). For the analysis of SCOTUS, although various corpora exist, each corpus consists of a limited fraction of the court resources and analyzes a specific perspective of the court, e.g., topic prediction on court opinions (Chalkidis et al., 2022a,b).

The LexGLUE corpus (Chalkidis et al., 2022a) is a benchmark dataset for legal language understanding and formulates the SCOTUS resource as a topic classification task. Specifically, the SCOTUS partition includes the opinion of the case,[3] and annotates the opinions with the issue area (topic), obtained from the Supreme Court Database (SCDB) (Spaeth et al., 2021), described in Section 3. Furthermore, Chalkidis et al. (2022b) considered the same SCOTUS task but with a focus on the fairness of classification models. They investigated various debiasing methods in the context of the legal domain. However, their fairness study of SCOTUS was limited to only one resource: opinion text.

To enrich the utilization of conversational data, expanding from Danescu-Niculescu-Mizil et al. (2012), Convokit (Chang et al., 2020) contains the transcripts of oral arguments of SCOTUS cases between 1955 to 2019.[4] This partition is also included as one of the sources in our Super-SCOTUS dataset (Section 3).

The SC-stance dataset (Bergam et al., 2022) combines the SCOTUS case questions annotated by the Oyez website and the corresponding SCOTUS opinion from a Kaggle dataset (Fiddler, 2022) to predict the political stands in court. They derived and labeled the political stands of the questions and opinions based on the winning side of cases from SCDB, i,e., favoring petitioners or respondents.

Henderson et al. (2022) proposed a large legal corpus, called Pile of Law, with approx. 256G (growing) legal and administrative text which concentrates on the US legal system, with the aim to provide a comprehensive corpus for legal text without containing toxic or private content. This corpus also includes SCOTUS opinions and the transcripts of oral arguments.

Bauer et al. (2023) annotated a dataset of approx. 436K US court opinions, including Supreme and Federal Courts, with key passages and summaries. However, their dataset is not publicly available.

As we discussed, current research only focuses on a limited partition of the SCOTUS resources. A more comprehensive dataset that covers various court resources is needed in order to understand how various court materials from different phases contribute to and reflect the Supreme Court. To address this gap, we provide a publicly-available multi-sourced SCOTUS dataset, connecting vari-

[2]We do not cover documents prior to hearing as these are not typically released to the public.

[3]Obtained from https://www.courtlistener.com/

[4]Obtained from https://www.oyez.org/

ous resources in different phases of the court procedure and enriching the dataset with additional annotations from the post-hearing phase.

3 The Super-SCOTUS Dataset

As shown in Figure 1, the court system comprises several stages. Since most pre-hearing documents are not publicly available, our dataset includes the resources from the court hearing and post-hearing, listed in Section 3.1. We detail our dataset construction in Section 3.2.

3.1 Dataset Sources

Supreme Court DataBase (SCDB)[5] This database (Spaeth et al., 2021) is recognized as the definitive source for analysis of the SCOTUS. It provides comprehensive structural metadata in over 50 categories such as justice votes, issue areas, and decision directions, for all SCOTUS cases between 1791 and 2021. Each case is labeled with a unique case ID.

Oral Arguments[6] We consider the oral arguments in the courtroom from Convokit (Chang et al., 2020), containing the transcripts of SCOTUS oral arguments among justices, advocates and amicus curiae for cases from 1955 to 2019. Each utterance is annotated with the speaker and the speaking time. Case IDs of the transcripts are also provided and aligned with the SCDB. It is worth noting that one case could have multiple conversations as cases can be re-argued at a later time.

Post-hearing Documents Justia[7] provides the following post-hearing data from 1791 to present: (1) **Syllabus**, a preliminary section of a court ruling that outlines the core facts and issues of the case, and the path that the case has been taken to the court, and (2) **Opinion**, set out the Court's judgment decision and its reasoning. The opinion might also include a *Concurrence* by judges who agree with the majority opinion but publish their own reasoning; and a *Dissent*, which provides the reasoning of justices who voted against the majority. Additionally, Justia provides (3) **Primary Holding**, a high-level summary of the case ruling, and (4) **Justia Summary**, a lay-friendly summary of the opinion. Primary holding and justice summary

	# cases covered
ConvoKit arguments	6,733
SCDB Metedata	6,721
Wikipedia Summary	1,191
Justia Syllabus	6,604
Justia Opinion	6,647
Justia Primary Holding	999
Justia Summary	602
Oyez Facts of the Case	2,945
Oyez Questions	2,946
Oyez Conclusions	2,944
Year Range	1955-2019
Justices	35

Table 1: Super-SCOTUS dataset statistics.

are additional case summaries, created by licensed attorneys from the platform with the aim to readability. This contrasts with the syllabus which was created by the court's reporter, and serves the primary purpose of official documentation.

Case Facts, Questions, and Conclusions For each SCOTUS case since 1789, Oyez[8] provides (1) **Facts of the Case**, summarizing the background of the case and the ruling from the lower court, (2) **Key Questions** addressed in the case, and (3) the **Conclusions**, i.e., answers to the key questions, also serving as a high-level summary of the voting reasoning. All information is collated by the editorial team of Justia.

Wikipedia Summary[9] Wikipedia provides a one-sentence summary of a set of notable Supreme Court cases from 1789 up to the present.

Justice Information We manually collect metadata for all justices from Wikipedia, Oyez, and the SCOTUS website,[10] with 10 attributes: year of birth; state of birth; gender; nominating president; nominating party; year of joining the court; chief justice (binary); year becoming chief justice; self-reported affiliated party; and ideal point (conservative or liberal) from Martin-Quinn score (MQ score; (Martin and Quinn, 2002)).

3.2 Dataset Creation

We crawled the data from the listed resources and automatically linked them at the case level. The overall statistics and coverage of SCOTUS cases

[5] http://scdb.wustl.edu/. A full list of SCDB attributes is provided in Appendix B.

[6] https://convokit.cornell.edu/documentation/supreme.html

[7] https://supreme.justia.com/

[8] https://www.oyez.org/

[9] https://en.wikipedia.org/wiki/Lists_of_United_States_Supreme_Court_cases, categorized by Chief Justice.

[10] https://www.supremecourt.gov/about/members_text.aspx

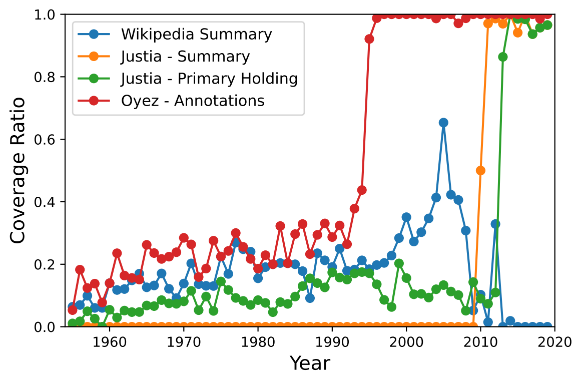


Figure 2: Case coverage of annotations. We show the annotations from Oyez (i.e., Facts of the Case, Question, and Conclusions) as "Oyez - Annotations" as the temporal coverages of those annotations are almost identical.

are shown in Table 1.[11] We used the Oral Arguments from ConvoKit as the base source, and augmented it with the other resources. In other words, Supreme Court cases that are not in the Convokit dataset are not included.

We linked the oral arguments with the SCDB based on the unique SCOTUS docket IDs. To parse the data from Justia, we separated the syllabus and opinion identifying the boundary of the two sections with a rule-based method. We separated the sections of Opinions, Concurrence, and Dissent based on the HTML tags provided by the website.[12] We automatically connected the Justica resources to the SCBD using case citations and discarded the Justia resources that are not explicitly matched by the citations. We linked cases in Oyez using the URLs in the oral argument data from Convokit. We utilized the case titles to match the data from the Wikipedia summary and the SCOTUS cases.

As shown in Table 1, the summaries from Justia, Oyez, and Wikipedia do not cover the SCOTUS cases comprehensively as they are not part of the official court-released documents. Annotations from Oyez and Justia are most complete for cases since 1995 and 2013, respectively (Figure 2). Wikipedia summaries exist only for notable cases, so no temporal pattern is observed.

[11]Detailed statistics of all resources are shown in Appendix A.

[12]Note that only 1109 (out of 6733) cases contain section separation tags. Sections could be also embedded in one whole document, i.e., no explicit HTML tags. We do not make separations for those cases.

Total Cases		5,205	
Oral Arguments	- avg. sentences - avg. words	462 10,866	
Syllabus	- avg. sentences - avg. words	33 837	
Decision Direction	- Conservative - Liberal	2,617 2,588	(50.3%) (49.7%)

Table 2: Statistics of the Decision Direction Prediction dataset, derived from Super-SCOTUS.

4 Decision Direction Prediction from Diverse Sources

As described in Section 2, most NLP research focuses on only a single type of document (predominantly Opinion documents (Zheng et al., 2021)) and typically a single task, e.g., the classification of the overall case topic (Chalkidis et al., 2022a,b). Here, we (a) propose a novel task of *decision direction prediction* which encodes information about case topic, justices' voting, and political leaning; and (b) consider this task across documents from the 'Court Hearing' and the 'Post Hearing' phase. Finally, we leverage the rich meta-data of Super-SCOTUS to study biases in the respective sources.

To demonstrate the utility of Super-SCOTUS, we study how different court materials reflect the *decision direction* of a case. Decision direction is coded as a binary variable as *Conservative* or *Liberal* by legal experts for the SCDB. Both the case outcome and the issue area (case topic) are taken into account during this labeling. For instance, in the issue area of criminal procedure, a Liberal label could indicate that the case outcome is 'pro civil liberties' or 'pro-underdog', and vice versa for Conservative. For cases related to economy, a Liberal label could indicate 'pro administrative action.[13]

4.1 Method

Data We consider two types of Super-SCOTUS documents, namely the Oral Arguments from the court hearing phase (OA), and the Syllabus from post-hearing (SL), and predict case decision direction. We remove cases which (i) do not have labels for voting result, decision direction, and issue area; (ii) were discussed in more than one oral argument session; (iii) lack the syllabus or opinion; or (iv) include companion cases, where multiple cases are jointly discussed. The statistics of the result-

[13]For full guidelines, see `http://scdb.wustl.edu/documentation.php?var=decisionDirection`.

ing dataset is shown in Table 2. Following previous work (Chalkidis et al., 2022a,b), the dataset is chronologically split into training (4133 cases, 1955-2002), development (536 cases, 2003-2010), and test (536 cases, 2011-2019) sets.

Protected attributes To further understand the model performance under different court documents, we investigate group disparities based on three groups which represent important perspectives of the cases: (1) **Issue Area**, indicating the general topic of the case (e.g., civil rights); (2) **Winning Side**, binary, whether the voting result favored the petitioner or respondent; and (3) **Vote Distribution**, the percentage of justices that vote with the majority categorized into five equal-width bins, indicating level of justice agreement.[14]

Models We include the following models: (1) BERT[15] (Devlin et al., 2019), pretrained on large generic domain corpus and served as a benchmark transformer-based model for various tasks, and (2) Legal-BERT[16] (Chalkidis et al., 2020), a domain-specific language model pretrained on English legal text, including court cases and legislation. We include the non-neural models Logistic Regression (LR) and Linear Support Vector Machine (SVM) with uni-gram TF-IDF features. We also report majority and random baselines.

For the non-neural baselines, we use grid search for the regularization parameter C with L2 penalty. For transformer-based methods, due to the input limitation, we consider either the first ($*$-first) or the last ($*$-last) 510 tokens of a particular document type, as the model input. We finetune models for 50 epochs with a batch size of 8, and a learning rate of 1e-5. We select the models with the best macro F1 on the development set.

4.2 Main Results

To study how different court materials reflect the case decision direction, we test how reliably decision direction can be predicted by various models based on different input documents. Table 3 shows that all models outperform the random and majority baseline, and that models based on the syllabus (bottom) are more reliable than oral argument-based prediction (center). This is unsurprising given that the syllabus is written after the court decision and explicitly states the case decision outcome, while oral arguments precede the decision and reflect justices' leanings implicitly at most as justices should not have made up their mind or aim to appear objective (Black et al., 2011; Dietrich et al., 2019). Holding the input type fixed, BERT outperforms Legal-BERT on our task. One possible reason is that the syllabus and oral arguments mostly contain generic words rather than legal vocabulary. Models trained on oral arguments perform better when using the first 510 tokens, which encode the case background and introduction. Models trained on the syllabus perform better when using the last 510 tokens, where the case decision is stated in terms, e.g., *affirmed* or *reversed* (but importantly *not* in terms of the ideological direction we are predicting here). We overall observe a relatively modest improvement of all models over the baseline, indicating the challenging nature of the task in particular for oral arguments-based models.

Input	Model	μ-F1	m-F1
-	Random	50.1	50.0
	Majority	52.1	32.4
Oral Argument	LR	52.4	50.1
	SVM	52.1	49.2
	BERT - first	**56.3**	**56.3**
	BERT - last	52.1	51.1
	Legal BERT - first	54.3	54.0
	Legal BERT - last	47.2	47.0
Syllabus	LR	57.1	57.1
	SVM	57.3	57.3
	BERT - first	53.7	53.7
	BERT - last	**66.2**	**65.9**
	Legal BERT - first	53.2	51.7
	Legal BERT - last	64.4	63.6

Table 3: Test results on decision direction prediction tasks with different inputs, i.e. oral arguments and syllabus. "μ-F1" and "m-F1" denote micro-F1 and macro-F1, respecitvely.

4.3 Group Disparity Analysis

Perhaps more interestingly than raw performance, we are now in a position to investigate different biases exhibited by models trained on different court documents. To this end, we analyze model performance for different *groups* of cases. We consider only the best (bolded) models from Table 3 based on the syllabus (BERT-last) and oral arguments (BERT-first). Following Chalkidis et al. (2022b), we first consider our three protected groups (Section 4.1) in isolation (Single Attribute) and subse-

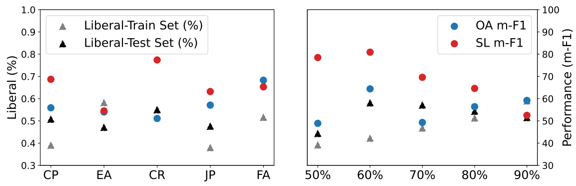
[14] We do not use vote count as not all cases are voted by all 9 justices.

[15] https://huggingface.co/bert-base-uncased

[16] https://huggingface.co/nlpaueb/legal-bert-base-uncased

Group	m-F1(↑)		train(%)(↑)	KL(↓)
	OA	SL		
Issue Area				
Criminal	55.9	68.7	**980 (24%)**	0.03
Economic	54.0	54.5	766 (19%)	0.02
Civil Rights	51.1	**77.4**	701 (17%)	**0.00**
Judicial Power	57.1	63.2	516 (12%)	0.02
1st Amendment	**68.3**	65.3	341 (8%)	0.05
Winning Side				
Petitioner	55.2	**68.1**	**2630 (64%)**	**0.00**
Respondent	**58.2**	60.8	1503 (36%)	**0.00**
Vote Distribution (% majority)				
50% - 60%	48.9	78.5	694 (17%)	0.01
60% - 70%	**64.4**	**80.9**	760 (18%)	0.05
70% - 80%	49.3	69.6	590 (14%)	0.02
80% - 90%	56.4	64.6	466 (11%)	**0.00**
90% - 100%	59.2	52.5	**1585 (38%)**	0.01

Table 4: Left: Group disparity results (mF1) across attribute groups with syllabus input for the best BERT models trained on Oral Arguments (OA) and on Syllabus (SL). Right: Representation Inequality ("train(%)") and Temporal Concept Drift ("KL"). The best (least harmful) values are highlighted in **bold**.



(a) Issue Area (b) Vote Distribution

Figure 3: Percentage of Liberal decisions in the train (grey) and test set (black) vs. Model Performance based on OA (blue) and SL (red) on the test set. "CP", "EA", "CR", "JP", and "FA" denote "Criminal Procedure", "Economic Activity", "Civil Rights", "Judicial Power" and "First Amendment", respectively. "k%" denotes Vote Distribution in $k\%$ - $(k+10)\%$, where $k \in [50, 60, 70, 80, 90]$.

quently their intersections (Cross-Attribute).

Single Attribute Disparity Table 4 (left) displays the macro-F1 score achieved for different partitions of cases, i.e., by Issue Area (top), Winning Side (center), and Vote Distribution (bottom). Overall, performance in different groups varies, especially in the groups of Issue Area and Vote Distribution with mF1 varying by > 10 points. Interestingly, for Vote Distribution the Syllabus-based model performs better when the voting pattern is more disparate, i.e., when the percentage of majority voting is close to 50%.

Additionally, although the SL-based model generally performs better than the OA-based model, there are select subgroups for which oral arguments may provide more signals on decision direction, e.g., 1st Amendment under Issue Area cases and high vote agreement (90%-100%) under Vote Distribution, indicating the difference of bias entailed in different court documents.

To rule out that performance disparity is caused by idiosyncracies of our data set, we consider two more general factors (Chalkidis et al., 2022b): (1) representation inequality, as number of training instances per group, and (2) Temporal Concept Drift as measured by the KL-divergence of per group label distribution between the training and test sets.

Table 4 (column #train(%)) shows the group representation in the training data, and reveals that performance does not correlate with representation. For instance, the largest Issue Area of "Criminal Law" does not achieve best performance. Turning to temporal concept drift (column KL), we observe low drift throughout. Drift does not predict group disparities either. For instance, for Issue Area the best performing labels from Syllabus and the Oral Arguments-based model have the highest and lowest KL, respectively.

Finally, we inspect if the label distribution correlates with group-specific model performance, focussing on the multi-class variables Issue Area and Vote Distribution. We show the label distributions of decision direction under different attribute groups, unpacking the KL-divergence in the train (grey) and test (black) sets, and compare them with the model performance in Figure 3. We do not observe severe label imbalance in the train and test sets across different attribute groups (little variance in the grey and black lines in Figure 3 (a) and (b)). We do not observe a significant correlation between Syllabus performance (SL, red) and label distribution (grey and black) under Pearson R ($p > 0.1$) across groups. The same holds for the relation of Oral Argument input (OA, blue) to label distribution.

Cross-Attribute Disparity Attributes are not necessarily independent of each other (Chalkidis et al., 2022b). To further unpack how different groups interact, we inspect performance discrepancies in groups that intersect on two of our attributes. In particular we intersect Issue Area with Winning

Side (Table 5 and 6, top) and Vote Distributions (Table 5 and 6, bottom). We do so for the OA-based model in Table 5 and the SL-based model in Table 6.

Table 5 (top) reveals for the OA-based model that, although models predicted 'Respondent Winners' with higher Macro-F1 in the overall analysis (Table 4), the picture is less clear in the intersection with Issue Area, where 'Petitioner Winners' are predicted with higher Macro-F1 for three out of five cases (CP, EA and most strikingly FA). Moving to the SL-based model (Table 6, top), we see that 'Petitioner Winners' are predicted with higher Macro-F1 than 'Respondent Winners' across issue areas, in agreement with the overall results in Table 4, suggesting the prevalent bias of Winning Side in the SL-based model. One possible reason is the explicit statement of the case outcome, which makes it easier for the SL-based model to learn the case direction signals. Further investigation, e.g., removing the voting result from the last few sentences in Syllabus (Malik et al., 2021), would be worthy to explore to understand the impact.

In Table 5, the performance varies in the cross-attributes of the Issue Area and Vote Distribution, indicating the complexity of identifying case direction across attributes from Oral Arguments. For our SL-based model in Table 6 (bottom), we observe that attributes in Issue Area that showed overall high performance, consistently achieve better performance across different Vote Distributions. Aligned with the overall results in Table 4, we also observe poor performance under unanimous cases throughout issue areas. One possible reason is that unanimous cases generally include important cases and historically those critical cases do not reflect much political leaning (Devins and Baum, 2017), one factor that would affect the case direction.

Overall, we show that models with different court documents as input (Oral Arguments vs. Syllabus) exhibit different performance discrepancies for different attribute groups. This is certainly due to the different nature in style and content of the documents. Our corpus and framework allow to systematically analyse downstream performance biases implied by different inputs – a mandatory prerequisite for any predictive models in high-stakes applications like the legal domain.

	Group A: Issue Area				
Group B:	CP	EA	CR	JP	FA
Petitioner	**52.4**	**59.2**	<u>46.4</u>	<u>54.1</u>	**79.2**
Respondent	<u>40.3</u>	<u>42.6</u>	**59.0**	**59.6**	<u>45.1</u>
50% - 60%	52.4	**60.6**	47.5	<u>26.7</u>	–
60% - 70%	**72.0**	<u>45.0</u>	**64.9**	**75.0**	–
70% - 80%	<u>44.0</u>	52.1	<u>24.5</u>	–	**73.3**
80% - 90%	49.7	49.7	–	–	–
90% - 100%	44.6	55.4	60.0	53.9	<u>66.7</u>

Table 5: MacroF1 results in cross-attribute influence on the BERT model with **Oral Arguments** input, intersecting Issue Area with either Winning Side (top) or Vote Distribution (bottom). **Best** and <u>worst</u> performing group per Issue Area are highlighted.

5 Conclusion and Future Vision

We presented a publicly-available multi-sourced dataset for the analysis of the US Supreme Court, namely the Super-SCOTUS dataset. Specifically, we focus on various resources from the phases of the court hearing and post hearing, connecting court-released data, both text documents (Opinions, Syllabus) and structured data (voting outcomes, issue areas), and enriching it with various summaries from different legal platforms, including key questions/conclusions of the court from Oyez and landmark case summaries from Wikipedia.

Our dataset supports a variety novel of NLP tasks (Section 5.1), where the empirical experiments in this paper only scratched the surface. To demonstrate the utility of Super-SCOTUS, we propose the challenging novel task of Decision Direction prediction, where labels encode the ideological direction most aligned with the case outcome (Liberal vs Conservative). We presented a case study using different court materials as input, namely Oral Arguments and Syllabus, and analyzed performance discrepancies of the best resulting models regarding three attribute groups. We observed that model performance varies with different inputs. The group disparity analysis further shows the performance difference across attributes, indicating that different biases exist in models trained on different court documents and suggesting the importance of considering various sources in the analysis of the court.

5.1 Future Vision

With the rich resources incorporated in our Super-SCOTUS corpus, we now have the opportunity to further enhance and expand the legal research in

	Group A: Issue Area				
Group B:	CP	EA	CR	JP	FA
Petitioner	**67.5**	**54.6**	**80.6**	**66.8**	**74.9**
Respondent	49.3	54.3	69.3	49.5	33.3
50% - 60%	**96.3**	44.8	81.5	**80.4**	–
60% - 70%	75.6	**71.8**	**100.0**	75.0	–
70% - 80%	66.4	55.6	**100.0**	–	**66.7**
80% - 90%	68.9	49.7	–	–	–
90% - 100%	45.2	52.1	61.0	43.3	56.4

Table 6: Same as Table 5, but with **Syllabus** based model input.

various tasks.

Court Decision Direction Prediction Our case study on decision prediction (Section 4) shows that different court documents—Oral Arguments versus Syllabus—entail different biases towards the prediction task. This analysis can be further expanded to other documents, e.g., Justia summary and Oyez conclusions, in order to gain a fuller understanding of court materials. The specific task of decision direction prediction would not be the end goal itself, but rather serve as vehicle to explore the biases and information encoded in those different documents, and how they impact downstream task performance. Furthermore, one might explore the impact of selecting certain information from those documents, or *summarizing* them automatically, would impact model performance.

Court Judgment Prediction We acknowledge the ethical concerns of legal judgement prediction, and do not recommend the task as benchmark task. That said, our data set allows to systematically study the extent to which justices' decisions are encoded in various court documents – most interestingly both during the court hearing (e.g., in Oral Arguments (Dietrich et al., 2019; Epstein et al., 2010; Epstein and Weinshall, 2021)) and post hearing. Related research casts the Court Judgment Prediction task based on documents from the post-hearing phase (Cui et al., 2022; Sim et al., 2015, 2016; Zhong et al., 2018) or purely on case characteristic (Ruger et al., 2004; Katz et al., 2017) and analyzes the results separately.

Ideology and Partisanship Most research develops the analysis of ideology and partisanship based on justices' voting (Martin and Quinn, 2002, 2007; Bailey and Maltzman, 2011; Bailey, 2013; Devins and Baum, 2017), assuming judicial preferences could be represented by their voting. It is also well-established that politicians choose words carefully in order to convey specific messages (Entman, 1993; Lakoff, 2010; Robinson et al., 2017). Building on our dataset, we now can investigate how the court documents, e.g., speeches from justices, advocates, or amicus curiae, reveal their partisan affiliations, and in how far their words align with their voting results (Bergam et al., 2022; Fang et al., 2023b).

Court Summarization Given the substantial length of most official legal documents and their high degree of expert language that is barely understandable to lay people, recent research has been focusing on extracting or summarizing key information from legal text (Ye et al., 2018; Wu et al., 2020; Shukla et al., 2022; Deroy et al., 2023; Bhattacharya et al., 2019; Shukla et al., 2022; Bauer et al., 2023).

Super-SCOTUS is a multi-reference summarization data set. It combines full-length court documents with various levels of summarization, aligned by case IDs. One promising and under-explored direction is to extract key contents from written Opinions and to generate the Syllabus. Additionally, the annotation of case questions, stating the key question addressed in the case, could be viewed as an 'extreme summary' of the case. Generating the key questions given the syllabus, opinion, and/or transcripts of oral arguments would be both challenging and a practically highly useful NLP application to increase accessibility of SCOTUS data to the general public. One natural related task would be how to answer those key questions, i.e., to generate the Conclusions in our data set, addressing the questions with judgment output and reasoning, given the facts of the case and other additional documents.

Model Speaker's Behaviors Increasing attention has been paid to understand the conversational behavior in the courtroom, e.g., association with voting (Epstein et al., 2010; Bergam et al., 2022; Epstein and Weinshall, 2021; Dietrich et al., 2019) and social dynamic in speakers' responses (Danescu-Niculescu-Mizil et al., 2012; Fang et al., 2023a). Particular for justices, with the rich metadata in our Super-SCOTUS annotations and various types of text from justices, we could further unlock research in analyzing court behaviors from different parties, e.g., to what extent the spoken text in the courtroom reveals the attributes of speakers, what linguist signals in spoken text those attribute groups encode, and furthermore how those signals correlate to their written documents.

Fairness and Bias In the context of the legal domain, one of the principal values is equality and non-discrimination (Xenidis and Senden, 2019). Although the definition of equality in law is completely the same, this also applies to the development of legal AI systems (Barfield, 2020; Chalkidis et al., 2022b; Wachter et al., 2020; Zhong et al., 2020). Particulary, Chalkidis et al. (2022b) benchmarked the evaluation of fairness in legal classification tasks over four jurisdictions, including SCOTUS with a task of issue area prediction along with two attribute groups (Respondent Type and Decision Direction). However, their analysis only focused on one specific court document, i.e., Opinions. With the availability of rich metadata (over 50 structural labels) and various text documents in phases of court hearing and post hearing, our Super-SCOTUS corpus provides an opportunity to systematically and comprehensively evaluate model fairness and debiasing methods in the classification tasks of SCOTUS.

Furthermore, attempts have also been made in summarizing or generating key content from written legal documents (Ye et al., 2018; Wu et al., 2020; Shukla et al., 2022; Bhattacharya et al., 2019). However, most research only focuses on the generation performance but not through the lens of fairness. Our multi-sourced dataset includes diverse layers of summaries connected with comprehensive metadata of the SCOTUS cases. This provides a unique opportunity to evaluate the fairness and bias of generation summarization models in the context of the legal domain and further benefit the development of fair legal systems.

Our creation of Super-SCOTUS takes a step further in this direction by giving the accessibility of comparing external resources with various SCOTUS-related materials, for instance, investigating partisanship alignment in the congressional records and various Supreme Court documents.

Ethics Statement

Social Impact We created a multi-sourced dataset which comprises diverse perspectives on the court hearing and post-hearing procedures, including both text and rich meta data. We presented a case study to showcase the potentials of our data set and discussed opportunities for future work. As a single, comprehensive resource Super-SCOTUS facilitates research in the legal domain from various perspectives (Section 5.1).

Acknowledging the importance of ethics in the legal domain (Tsarapatsanis and Aletras, 2021), we follow Chalkidis et al. (2022b) in arguing that the development of legal technology should not only rely on the performance of the majority group, and add that it additionally should not rely on only a single document source. Our data set allows to study these questions, e.g., by identifying the confounders in court documents that lead to unfair or unreliable behavior.

Personal Information Super-SCOTUS contains personal information, e.g., about the petitioners and respondents (from the SCDB data base). Given that all cases are (reasonably) high profile by definition of being considered at the SCOTUS, and the data was obtained from public repositories we did not redact this information. We encourage researchers to consider the potential impacts of personal information in their research. The usage of the data has to be in compliance with US law.

Additionally, we manually added gender and nominated party of SCOTUS justices as additional meta data, again sourced from public databases (Wikipedia). Given the impact of Supreme Court justices on the US legal system, we collect this data to provide more comprehensive context for the analysis of the US Supreme Court as in the whole system, but not at the individual level. We make no attempt to target any justices, nor encourage it.

Credit Attribution / Licensing The Convokit dataset is distributed under the MIT license. The usae of official data by the Supreme Court (SCDB) must comply with US law. Annotations from Justia are provided as open access contributed pro-bono by licensed attorneys.[17] Data from Oyez is under Creative Commons Attribution-NonCommercial 4.0 International License.[18] Annotations from Wikipedia are available under the Creative Commons Attribution-ShareAlike License 4.0.[19] We release the Super-SCOTUS dataset under a CC-BY-NC-SA-4.0 license.[20]

[17] https://law.justia.com/annotations/

[18] https://creativecommons.org/licenses/by-nc/4.0/

[19] https://en.wikipedia.org/wiki/Wikipedia:Text_of_the_Creative_Commons_Attribution-ShareAlike_4.0_International_License

[20] https://creativecommons.org/licenses/by-nc-sa/4.0/

References

Michael A Bailey. 2013. Is today's court the most conservative in sixty years? challenges and opportunities in measuring judicial preferences. *The Journal of Politics*, 75(3):821–834.

Michael A Bailey and Forrest Maltzman. 2011. *The constrained court: Law, politics, and the decisions justices make*. Princeton University Press.

Woodrow Barfield. 2020. *The Cambridge Handbook of the Law of Algorithms*. Cambridge University Press.

Emmanuel Bauer, Dominik Stammbach, Nianlong Gu, and Elliott Ash. 2023. Legal extractive summarization of us court opinions. *arXiv preprint arXiv:2305.08428*.

Noah Bergam, Emily Allaway, and Kathleen Mckeown. 2022. Legal and political stance detection of SCOTUS language. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 265–275, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Paheli Bhattacharya, Kaustubh Hiware, Subham Rajgaria, Nilay Pochhi, Kripabandhu Ghosh, and Saptarshi Ghosh. 2019. A comparative study of summarization algorithms applied to legal case judgments. In *Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part I 41*, pages 413–428. Springer.

Ryan C Black, Sarah A Treul, Timothy R Johnson, and Jerry Goldman. 2011. Emotions, oral arguments, and supreme court decision making. *The Journal of Politics*, 73(2):572–581.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022a. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Schwemer, and Anders Søgaard. 2022b. FairLex: A multilingual benchmark for evaluating fairness in legal text processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4389–4406, Dublin, Ireland. Association for Computational Linguistics.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. ConvoKit: A toolkit for the analysis of conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60, 1st virtual meeting. Association for Computational Linguistics.

Junyun Cui, Xiaoyu Shen, Feiping Nie, Zheng Wang, Jinglong Wang, and Yulong Chen. 2022. A survey on legal judgment prediction: Datasets, metrics, models and challenges. *arXiv preprint arXiv:2204.04859*.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, pages 699–708.

Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. 2023. How ready are pre-trained abstractive models and llms for legal case judgement summarization? *arXiv preprint arXiv:2306.01248*.

Neal Devins and Lawrence Baum. 2017. Split definitive: How party polarization turned the Supreme Court into a partisan court. *The Supreme Court Review*, 2016(1):301–365.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bryce J Dietrich, Ryan D Enos, and Maya Sen. 2019. Emotional arousal predicts voting on the us supreme court. *Political Analysis*, 27(2):237–243.

Robert M Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of communication*, 43(4):51–58.

Lee Epstein, William M Landes, and Richard A Posner. 2010. Inferring the winning party in the Supreme Court from the pattern of questioning at oral argument. *The Journal of Legal Studies*, 39(2):433–467.

Lee Epstein and Keren Weinshall. 2021. *The Strategic Analysis of Judicial Behavior: A Comparative Perspective*. Cambridge University Press.

Biaoyan Fang, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2023a. It's not only What You Say, It's also Who It's Said to: Counterfactual Analysis of Interactive Behavior in the Courtroom. In *Proceedings of The 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, Nusa Dua, Bali. Association for Computational Linguistics.

Biaoyan Fang, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2023b. More than Votes? Voting and Language based Partisanship in the US Supreme Court. In *Findings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore. Association for Computational Linguistics.

Garrett Fiddler. 2022. SCOTUS Opinions. Full text and metadata of all opinions written by SCOTUS justices through 2020.

Peter Henderson, Mark Krass, Lucia Zheng, Neel Guha, Christopher D Manning, Dan Jurafsky, and Daniel Ho. 2022. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. In *Advances in Neural Information Processing Systems*, volume 35, pages 29217–29234. Curran Associates, Inc.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. CUAD: An expert-annotated nlp dataset for legal contract review. *NeurIPS*.

Daniel Martin Katz, Michael J Bommarito, and Josh Blackman. 2017. A general approach for predicting the behavior of the supreme court of the united states. *PloS one*, 12(4):e0174698.

George Lakoff. 2010. *Moral politics: How liberals and conservatives think*. University of Chicago Press.

Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. 2019. Claudette: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27:117–139.

Daniel Locke and Guido Zuccon. 2018. A test collection for evaluating legal case law search. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 1261–1264, New York, NY, USA. Association for Computing Machinery.

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4046–4062, Online. Association for Computational Linguistics.

Andrew D Martin and Kevin M Quinn. 2002. Dynamic ideal point estimation via Markov chain Monte Carlo for the US Supreme Court, 1953–1999. *Political analysis*, 10(2):134–153.

Andrew D Martin and Kevin M Quinn. 2007. Assessing preference change on the US Supreme Court. *The journal of law, economics, & organization*, 23(2):365–385.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 19–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael D Robinson, Ryan L Boyd, Adam K Fetterman, and Michelle R Persich. 2017. The mind versus the body in political (and nonpolitical) discourse: Linguistic evidence for an ideological signature in US politics. *Journal of Language and Social Psychology*, 36(4):438–461.

Theodore W Ruger, Pauline T Kim, Andrew D Martin, and Kevin M Quinn. 2004. The supreme court forecasting project: Legal and political science approaches to predicting supreme court decision making. *Columbia law review*, pages 1150–1210.

Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, and Saptarshi Ghosh. 2022. Legal case document summarization: Extractive and abstractive methods and their evaluation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1048–1064, Online only. Association for Computational Linguistics.

Yanchuan Sim, Bryan Routledge, and Noah Smith. 2015. The utility of text: the case of amicus briefs and the supreme court. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Yanchuan Sim, Bryan Routledge, and Noah A. Smith. 2016. Friends with motives: Using text to infer influence on SCOTUS. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1733, Austin, Texas. Association for Computational Linguistics.

Harold Spaeth, Lee Epstein, Ted Ruger, Jeffrey Segal, Andrew D. Martin, and Sara Benesh. 2021. Supreme court database, version 2021 release 01. Washington University Law.

Robert L Stern and Eugene Gressman. 1950. *Supreme Court practice: jurisdiction, procedure, arguing and briefing techniques, forms, statutes, rules for practice in the Supreme Court of the United States*. Bureau of national affairs.

Dimitrios Tsarapatsanis and Nikolaos Aletras. 2021. On the ethical limits of natural language processing on legal text. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3590–3599, Online. Association for Computational Linguistics.

Don Tuggener, Pius von Däniken, Thomas Peetz, and Mark Cieliebak. 2020. LEDGAR: A large-scale multi-label corpus for text classification of legal provisions in contracts. In *Proceedings of the Twelfth*

Language Resources and Evaluation Conference, pages 1235–1241, Marseille, France. European Language Resources Association.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2020. Bias preservation in machine learning: the legality of fairness metrics under eu non-discrimination law. *W. Va. L. Rev.*, 123:735.

Yiquan Wu, Kun Kuang, Yating Zhang, Xiaozhong Liu, Changlong Sun, Jun Xiao, Yueting Zhuang, Luo Si, and Fei Wu. 2020. De-biased court's view generation with causality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 763–780, Online. Association for Computational Linguistics.

Raphaële Xenidis and Linda Senden. 2019. Eu non-discrimination law in the era of artificial intelligence: Mapping the challenges of algorithmic discrimination. *Raphaële Xenidis and Linda Senden,'EU non-discrimination law in the era of artificial intelligence: Mapping the challenges of algorithmic discrimination'in Ulf Bernitz et al (eds), General Principles of EU law and the EU Digital Order (Kluwer Law International, 2020)*, pages 151–182.

Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. 2018. Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1854–1864, New Orleans, Louisiana. Association for Computational Linguistics.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, page 159–168, New York, NY, USA. Association for Computing Machinery.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549, Brussels, Belgium. Association for Computational Linguistics.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online. Association for Computational Linguistics.

A Detailed Dataset Statistic

Table 7 provides a detailed statistic of the Super-SCOTUS corpus.

B A Full list of SCDB Attributes

Table 8 provides a full list of SCDB attributes.

	# cases covered		Number
ConvoKit arguments	6,733	- avg. sentences - avg. words	512 12,191
SCDB Metedata	6,721	- # attributes	53
Wikipedia Summary	1,191	- avg. sentences - avg. words	1 71
Justia Syllabus	6,604	- avg. sentences - avg. words	32 791
Justia Opinion	6,647	- avg. sentences - avg. words	340 9,276
Justia Primary Holding	999	- avg. sentences - avg. words	1 37
Justia Summary	602	- avg. sentences - avg. words	12 305
Oyez Facts of the Case	2,945	- avg. sentences - avg. words	8 199
Oyez Questions	2,946	- avg. sentences - avg. words	1 36
Oyez Conclusions	2,944	- avg. sentences - avg. words	7 211

Table 7: Detailed statistics of Super-SCOTUS.

0	adminAction	1	adminActionState	2	authorityDecision1
3	authorityDecision2	4	caseDisposition	5	caseDispositionUnusual
6	caseId	7	caseIssuesId	8	caseName
9	caseOrigin	10	caseOriginState	11	caseSource
12	caseSourceState	13	certReason	14	chief
15	dateArgument	16	dateDecision	17	dateRearg
18	decisionDirection	19	decisionDirectionDissent	20	decisionType
21	declarationUncon	22	docket	23	docketId
24	issue	25	issueArea	26	jurisdiction
27	lawMinor	28	lawSupp	29	lawType
30	lcDisagreement	31	lcDisposition	32	lcDispositionDirection
33	ledCite	34	lexisCite	35	majOpinAssigner
36	majOpinWriter	37	majVotes	38	minVotes
39	naturalCourt	40	partyWinning	41	petitioner
42	petitionerState	43	precedentAlteration	44	respondent
45	respondentState	46	sctCite	47	splitVote
48	term	49	threeJudgeFdc	50	usCite
51	voteId	52	voteUnclear		

Table 8: A list of SCDB attributes (Spaeth et al., 2021).